A magyar Braille-rövidírás megújítása félautomatikus módszerrel

Sass Bálint

MTA Nyelvtudományi Intézet
sass.balint@nytud.mta.hu

Kivonat A dolgozatban az új magyar Braille-rövidírást illetve létrehozásának módját mutatjuk be. A félautomatikus eljárás két részből áll: egy automatikus, korpuszvezérelt elven működő algoritmus határozza meg a legalkalmasabb rövidítendő elemeket; ezt követi a manuális véglegesítő lépés a kényelmes használhatóság szempontjainak figyelembevételével. A létrejött 33 elemű szabályrendszer könnyen megtanulható, jól olvasható, jól felismerhető. Rövidítési képessége 13,3%, mely 3,4%-kal növeli meg a ma használatos kis rövidírás (9,9%) hatékonyságát. Az új rövidírás alkalmas a vakok általi tesztelésre és majdani használatra.

Kulcsszavak: Braille-írás, Braille-rövidírás, korpuszvezérelt, rövidítés, rövidítési képesség, szabály, gyakoriság, használhatóság

1. Célkitűzés

A vakok által világszerte használt, tapintáson alapuló Braille-írásnak számos nyelvre létezik ún. Braille-*rövidírás* változata (angol: [1]; német: [2,3]). Ezek az általános Braille-írást nyelvspecifikus rövidítési, tömörítési szabályokkal egészítik ki. Rövidírás használatával gyorsul az írás-jegyzetelés és az olvasás folyamata. Napjainkban, a speciális Braille-nyomtatók egyre szélesebb körű elterjedésével az is fontos, hogy a rövidírással írt szöveg kinyomtatva jelentősen kisebb terjedelmű. 2012-2013-ban valósult meg a projekt a Magyar Vakok és Gyengénlátók Országos Szövetsége és az MTA Nyelvtudományi Intézet együttműködésében, melynek keretében a magyar Braille-rövidírást a mai nyelvhasználatot is figyelembe vevő új rövidítésekkel bővítjük, azzal a céllal, hogy a rövidítési képessége a korábbi nagyjából 10%-ról jelentős mértékben, akár 15-20% közelébe növekedjen [4].

2. A magyar Braille-írás

A Braille-karakterek (ún. Braille-cellák) két oszlopban elrendezett 3-3, azaz összesen hat kidomborodó pontból állnak. Az egyes pontokra a következő elrendezésben számokkal hivatkozunk: $\begin{smallmatrix}1&4\\2&5\\3&6\end{smallmatrix}$. A kidomborodó és ki nem domborodó pontok mintázataiból összesen $2^6 = 64$ féle különböző karakter áll elő. A tapintható írásrendszerek történeti bemutatásáról l. [5]-t.

2.1. A magyar teljesírás

A magyar Braille-karaktereket az 1. táblázatban láthatjuk a Görgényi Miklóstól átvett elrendezésben [6, 1. fejezet]. Az első sorban a felső 4 pontot elfoglaló jelek vannak, a következő 3 sor ehhez teszi hozzá rendre a 3-as, a 3-as/6-os, illetve a 6-os pontot. Az 5. sor az első sor jeleinek *csúsztatott*, azaz egy ponttal lefelé mozdított megfelelőit tartalmazza. Az utolsó három oszlopban vannak azok a jelek, melyekben nem szerepel az 1-es/2-es pont. A bal és jobb oszlopban is, valamint a felső és alsó sorban is pontot tartalmazó ún. *erős* (angolban: *strong* [1, 28. oldal]) jeleket szürke háttérrel jelöltük. Az $5 \cdot 13 = 65$ pozíción szépen ábrázolódik a 64 jel, az üres jel kétszer fordul elő (a 11. oszlopban). Az angol ábécé betűi nagyjából az első három sor első tíz oszlopában találhatók.

1. táblázat. A magyar Braille-ábécé.

1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
⠁ [a]	⠃ [b]	⠉ [c]	⠙ [d]	⠑ [e]	⠋ [f]	⠛ [g]	⠓ [h]	⠊ [i]	⠚ [j]	⠀ [␣]	⠡ [á]	⠈ [@]
⠅ [k]	⠇ [l]	⠍ [m]	⠝ [n]	⠕ [o]	⠏ [p]	⠪ [ö]	⠗ [r]	⠎ [s]	⠞ [t]	⠲ [.]	⠌ [í]	⠵ [Z]
⠥ [u]	⠧ [v]	⠭ [x]	⠽ [y]	⠵ [)]	⠟ [q]	⠼ [#]	⠳ [ü]	⠦ [(]	⠾ [ű]	⠤ [-]	⠬ [ú]	⠙ [D]
⠮ [é]	⠵ [z]	⠠ [C]	⠛ [G]	⠎ [S]	⠝ [N]	⠪ [ő]	⠞ [T]	⠬ [ó]	⠺ [w]	⠠ [V]	⠁ [A]	⠇ [L]
⠂ [,]	⠆ [;]	⠒ [:]	⠓ [H]	⠦ [?]	⠖ [!]	⠶ [=]	⠴ [”]	⠔ [*]	⠠ [Q]	⠀ [␣]	⠗ [R]	⠋ [F]

A Braille-ábécé nagybetűket nem tartalmaz, a táblázatban szereplő nagybetűk speciális jelentéssel bírnak. A tényleges nagybetűket két Braille-karakterrel írjuk le: a nagybetűjelölő ⠠ [A] után írjuk a megfelelő betűt, pl. *Ő* = ⠠⠪. A táblázatban tehát a speciális karakterek – amilyen a nagybetűjelölő is – átírására egyezményes nagybetűs jeleket használunk. Az eredeti táblázatot kicsit módosítva a magyar teljesírás újabb szabályainak megfelelően W helyett @-t, pont helyett H-t, X helyett pontot írtam. A magyar ékezetes betűk és kettősbetűk – legtöbbször tükrözéssel kialakított – önálló Braille-jellel rendelkeznek, utóbbiak a következő bevett jelölésekkel: cs=C, gy=G, ly=L, ny=N, sz=S, ty=T, zs=Z. További karakterek: az ⠈ [R] ékezetjelölő; a ⠨ [V] védő karakter, ha ez szerepel egy karakter(sorozat) előtt, akkor azt nem rövidítésként, hanem literálisan, karakterenként kell olvasni; a ⠙ [D] a számjelölő, mely az *a*-tól *j*-ig terjedő betűkből számjegyet képez (az 5 Braille-ben ⠙⠑ [De]); a ⠓ [H] és az ⠋ [F] jelhez pedig nincs jelentés rendelve, ezek ún. *szabad gyök*ök. A táblázatban látható karakterekkel lehet az egyes Braille-cellákat sima szövegre átírni, ezt az átírást a dolgozatban használni is fogjuk: a *Magyarország* szó átírása például *AmaGarorSág*.

2.2. A magyar „kis” rövidírás

Amint már utaltunk rá, a Braille-rövidírások az írott anyag tömörítésére szolgálnak, csökkentve a helyigényt és gyorsítva az írás-olvasást. A rövidírás rövidítési

szabályok gyűjteménye. Ma hazánkban sztenderd módon az ún. „kis” rövidírást használják, ami a korábbi jóval bonyolultabb nagy rövidírás [7] könnyen megjegyezhető szabályaiból áll. A kis rövidírás rendszerét a 2. táblázatban foglaltuk össze [6] 7. fejezete alapján. Feltüntettük a mért rövidítésiképesség-értékeket is.

2. táblázat. A kis rövidírás szabályai.

rövidítéscsoport	rövidítési képesség
1. Nagybetűjel ⁝ törlése.	+2,3%
2. Vessző utáni szóköz törlése.	+1,4%
3. A határozott névelők rövidítése: r(⁘[az])=⁝[.], r(⁘*[a]*)=⁖*[,]*), és az utánuk lévő szóköz törlése.	+1,9%
4. Az alábbi 44 szabály alkalmazása.	+4,3%

7 szóvégi rövidítés		16 egyjelű szórövidítés				21 kétjelű szórövidítés					
-ban/-ben	b	Cak	C	meL	m	aNNi	ai	mind	md	rövid	rd
-ból/-ből	b.	de	d	nem	n	boldog	bg	mint	mt	forr	rr
-hoz/-hez/-höz	h.	és	é	óta	ó	eNNi	ei	orSág	og	Sabad	Sd
ként	k.	hoG	h	pedig	p	gond	gd	olvas	os	tanáC	tC
-ról/-ről	r.	is	i	tehát	t	függ	gg	öSSe	öe	teljes	ts
-tól/-től	t.	íG	í	után	u	Gors	Gs	pont	pt	világ	vg
-val/-vel	v	kell	k	úG	ú	keres	ks	pénz	pz	volt	vt
		leS	l	van	v						
+1,4%		+1,8%				+1,1%					

A kis rövidírás rövidítési képesség mindösszesen: 9,9%

A *-val/-vel* ragot hasonulás esetén is *v*-vel rövidítjük. A pontot tartalmazó szóvégi rövidítéseknek bevezetés alatt áll egy újabb formája, mely a rag első és utolsó betűjéből áll. Az egyjelű szórövidítéseket önálló szóként és összetételben, a kétjelűeket ezen kívül bármilyen toldalékolt forma esetén is alkalmazzuk. Előfordul, hogy egy rövidítésként is értelmezhető karaktersort literálisan akarunk értelmeztetni, ilyenkor a már említett ⁝[V] védőjellel prefixáljuk a karaktersort. A ⁝⁘⁘⁝[Serb] tehát a *szer* főnév *-ban/-ben*-ragos alakja, a ⁝⁘⁘⁝⁝[SerVb] viszont a *szerb* népnév.

Látjuk, hogy az első 3 csoportban lévő „trükkös” (információvesztő és szóközlenyelő) szabályok nagyon jelentős rövidítési képességgel bírnak. A negyedik csoportban lévő szabályoknál az egy szabályra eső rövidítési képesség folyamatosan csökken, rendre: 0,2%, 0,12%, a kétjelű szórövidítéseknél – melyek sok esetben viszonylag ritka szavakat rövidítenek – pedig csak 0,05%.

3. Rövidítési képesség vs. használhatóság

Célunk tehát a magyar rövidírás rövidítési képességének jelentős növelése, de nem csak ez a szempont vezérli a szabályrendszer kialakítását. Ugyanilyen fontos

az is, hogy a végső rendszer kényelmesen használható legyen. A munkálatok során a következő használhatósági követelmények körvonalazódtak:

1. az új szabályrendszer az ismert kis rövidírást egészítse ki;
2. *jó olvashatóság*: a rövidítések emlékeztessenek az eredetire;
3. *jó felismerhetőség*: tapintás útján könnyen felismerhető jelek alkalmazása;
4. *könnyű megtanulhatóság*: kevés, egyszerű szabály.

A nagy rövidítési képesség és kényelmes használhatóság egymás ellen ható követelmények, itt egy körültekintően kidolgozott kompromisszumra van szükség annak érdekében, hogy a potenciális felhasználók elfogadják és szívesen alkalmazzák az új rövidírást. A fenti feltételeknek kevéssé megfelelő nagy rövidírás éppen bonyolultsága miatt nem terjedt el korábban. Jelen dolgozatban bemutatjuk a kialakított kompromisszumos javaslatot, mely törekszik a maximális rövidítési képesség elérésére, miközben megfelel a használhatósági feltételeknek is.

4. Alapötlet: korpuszvezéreltség

A rövidírás-rendszerek kifejlesztése sok esetben nagy időigényű feladat, az egységes angol rövidírás kialakítása majdnem két évtizedet vett igénybe [8].

Jelen munkálat alapötlete azon a felismerésen alapul, hogy az ideális rövidítési szabályok a magyar nyelv rendelkezésre álló korpuszgyakorisági adatai alapján, korpuszvezérelt módon, automatikusan meghatározhatók, és ezek alapján a lehető legnagyobb rövidítési képességgel bíró új magyar rövidírás záros határidőn belül elkészíthető. A háttérben az az egyszerű gondolat van, hogy nyilván a lehető leggyakoribb elemeket (betűsorozatokat) érdemes a lehető legrövidebbre rövidíteni, ekkor nyerjük összességében a legtöbbet. A rövidség szempontjából tehát nem volt ideális választás annak idején a ritka *ty* kettősbetű egyjelű rövidítése: r(⠽⠞[ty])=⠾[T] a nála akár 20× gyakoribb kétkarakteres elemek (pl.: *et*) helyett. A fenti gondolat kiegészül azzal, hogy „mohó" eljárást követünk, azaz mindig azt az aktuális új szabályt választjuk, ami éppen a legnagyobb rövidítést eredményezi.

Cél volt a lehető legkisebb emberi beavatkozás, de nyilvánvalóvá vált, hogy a használhatósági feltételeknek való megfelelés nehezen automatizálható. A teljes folyamat tehát nem megy automatikusan: szükséges a szakértői közreműködés a szabályok kézi véglegesítése során. Leegyszerűsítve mondhatjuk, hogy automatikusan áll elő az, hogy *mit* rövidítünk, és manuálisan, hogy *mire*.

A kutatás során a fenti ötlet szerint jártunk el, mert így minden említett követelménynek meg tudtunk felelni, és formailag is olyan szabályokat tudtunk alkotni, melyek hasonlóan a kis rövidírásban használt szabályokhoz. Alább néhány alternatív megközelítést említek. Szóba kerülhet (1) a gyorsírás vizsgálata; (2) az sms és/vagy twitter korpuszok vagy (3) rövidítéstárak tanulmányozása; vagy annak direkt felmérése, hogy (4) a fiatal vakok hogyan rövidítenek. Nem járható út (5) a teljes magánhangzó-elhagyás ötlete a nehéz olvashatóság, (6) prefixfák használata pedig az írás nehezítettsége miatt. Érdemes lehet megvizsgálni (7) az

általános tömörítő algoritmusok architektúráját is. Egy ilyen módszert alkalmazott Arató András kandidátusi értekezésében [9, 7. fejezet], melynek keretében a rövidírás automatikus kialakíthatóságát is érinti.

5. A rövidítendő elemeket megadó automatikus eljárás

Az alapelv tehát az, hogy a lehető leggyakoribb és leghosszabb elemeket próbáljuk a lehető legrövidebbre rövidíteni. Ezt úgy valósítjuk meg, hogy meghatározzuk az adott helyzetben éppen legnagyobb rövidítési képességet adó szabályt, majd az így kialakult új helyzetben ismét az aktuális legnagyobb rövidítési képességet adó szabályt, és így tovább. A rövidítési képességet – jele: rk() – az 1. ábrán látható módon számítjuk.

$$\mathrm{rk}(w, r(w)) = [l(w) - l(r(w))] \cdot \mathrm{fq}(w)$$

1. ábra. A rövidítési képesség számítása. w – az eredeti rövidítendő karaktersorozat, $r(w)$ – a rövidítés, $l()$ – a hossz (karakterszám), fq() – a gyakoriság (millió szóra eső előfordulási szám). Megjegyzés: az empirikus úton meghatározott, azaz korpuszon mért rövidítési képességre is ugyanúgy az rk() jelölést alkalmazzuk, a szövegösszefüggésből mindig világos, hogy a kettő közül melyikre gondolunk.

A *sem* elem gyakorisága (3302) például kicsivel több, mint a *Serint* elem gyakorisága (2515). Ha egy karakterre rövidítünk – például s és S a két rövidítésjelölt –, a rövidítési képesség a következő: $\mathrm{rk}(\mathrm{sem}) = (3-1) * 3302 = 6604$, valamint $\mathrm{rk}(\mathrm{Serint}) = (6-1) * 2515 = 12575$. Nyilvánvalóan érdemes a (ritkább) *Serint*-et rövidíteni, mivel így az elért rövidülés majdnem kétszeres. De még akkor is ugyanez lenne a jó döntés, ha a *sem* egyjelű és a *Serint* kétjelű rövidítése közül kellene választanunk: $\mathrm{rk}(\mathrm{Serint}) = (6-2) * 2515 = 10060$.

Az algoritmus a következő lépésekből áll.

1. Korpusz alapján elkészítjük a szavak gyakorisági listáját.
2. Számba vesszük az összes elvben rövidíthető nyelvi elemet: a szavak gyakorisági listájából származtatjuk a karakter-n-gramok gyakorisági listáját, minden karakter-n-gramot külön-külön számolva.
3. A rendelkezésre álló rövidítésjelek hosszának ismeretében kiszámoljuk az elemek rövidítési képességét (1. ábra).
4. Rendezzük a gyakorisági listát rövidítési képesség szerint.
5. A rendezett lista *első* bejegyzése adja a maximális rövidítési képességet, vesszük ezt az elemet, és hozzárendelünk egy megfelelő rövidítést.
6. Az így kialakított rövidítő szabályt *alkalmazzuk* a szavak gyakorisági listájára.

7. Újból a 2. pontra lépünk, amíg el nem érünk egy elfogadható összesített rövidítési képességet, vagy a szabályrendszer mérete át nem lép egy adott küszöböt.

Ez az algoritmus a korábban javasolt algoritmus [4] továbbfejlesztett változata, mely az éppen létrehozott szabálynak a szavak gyakorisági listájára való alkalmazása által jól kezeli a rövidítendő elemek esetleges átfedésének a rövidítésiképesség-értékekre gyakorolt módosító hatását. Azért szükséges mindig az eredeti, szavakat tartalmazó gyakorisági listára alkalmazni a szabályt, mert az n-gramok adott esetben egy rövidítendő elemnek csak egy részletét tartalmazzák az n-gram elején vagy végén.

Pontosan azért van szükség a rövidítési képességek újraszámolására, és a lista újrarendezésére, mert az éppen aktuálisan létrehozott szabály alkalmazása befolyásolja (csökkenti) bizonyos elemek hosszát, így az azok esetleges továbbrövidítése során elérhető rövidülés változik. Az újrarendezés során létrejövő sorrendváltozást, helycserét a 2. ábrán szemléltetjük.

w	$\mathrm{fq}(w) \cdot l(w) - l(r(w))$	$\mathrm{rk}(w)$
Ser	10901	$\times 1 = 10901$
Serint	2515	$\times 4 = 10060$
maGar	3326	$\times 3 = 9978$
...		

(a) A gyakorisági lista eleje.

w	$\mathrm{fq}(w) \cdot l(w) - l(r(w))$	$\mathrm{rk}(w)$
maGar	3326	$\times 3 = 9978$
...		
Srint	2515	$\times 3 = 7545$
...		
Sr	10901	$\times 0 = 0$

(b) Sorrend a *Ser* → *Sr* rövidítés után.

2. ábra. Sorrendváltozás a futás során.

A 2(a) ábrán karakter-n-gramok gyakorisági listájának eleje látható az algoritmus futásának egy pontján. Tegyük fel, hogy két karakterre rövidítünk. Az algoritmus 5. pontja szerint a rövidítendő elem a *Ser* lesz, a hozzá társított rövidítés legyen a *Sr*. Ennek az új szabálynak a szógyakorisági listára történő alkalmazása után (algoritmus 6., 7. és 2. pont) áll elő a 2(b) ábrán látható helyzet. Mivel a *Serint*-ből az új szabály az egy karakterrel rövidebb *Srint*-et hozza létre, ez az elem (jóval) lejjebb kerül a listán, és a következő rövidítendő elem a *maGar* lesz. A *Ser*-ből lett *Sr* kétjelű rövidítéssel nyilván nem rövidíthető tovább. Kiemelendő, hogy kialakulhat olyan helyzet, hogy az algoritmus további futása során a *Srint* végül mégiscsak a lista elejére kerül és (tovább)rövidül.

A következő példa is az algoritmus működését világítja meg. Egy karakterre rövidítve a legnagyobb rk()-gel rendelkező magyar karaktersorozat az *et*: ha *y*-nal rövidítjük, akkor a (főként igevégződésként) nagyon gyakori *ett* elem *yt*-ként jelenik meg. Ha azonban úgy döntünk, hogy csak kétjelű rövidítéseket alkalmazunk, akkor az *et* nem rövidül ($\mathrm{rk}() = 0$), viszont az algoritmus futása során hamar a lista elejére kerül az *ett*.

A rövidítendő elemek esetleges *átfedése* miatt nem lehet azt megtenni, hogy a szabályokat egymástól függetlenül alakítjuk ki. Az *et* egyjelű rövidítése önmagá-

ban rk() = 0, 77%-os, a *te* elemé pedig 0, 57%-os rövidülést jelent. A gyakorlatban egymás után alkalmazva a két szabályt már csak 0, 77% + 0, 24% az eredmény, mivel az első szabály nagyon sokszor az *ete* elem első két karakterét rövidíti, elvéve ezzel a lehetőséget a második szabály elől. Az épp kialakított új szabály teljes szógyakorisági listára való alkalmazása (algoritmus 6. pont) oldja meg ezt a problémát. Így mivel minden ponton világos, hogy adott szabálynak mennyi a tényleges rk()-e, könnyen kiválasztható a legjobb szabály, a szabályok egymással összefüggésben, egymásra épülve jönnek létre.

Olyan eset azonban előfordulhat, hogy egy szabály bal oldala egy rövidítést teljes egészében tartalmaz, ezt nevezzük *továbbrövidítés*nek. Ha például r(*ek*)=*!*, és a lista elejére kerül a (már csak 4 karakteres) *Ger!* elem, akkor ezt rövidíthetjük *Gk*-val. A továbbrövidítést kifejtve két független szabályt kapunk – r(*Gerek*)=*Gk* és r(*ek*)=*!* –, melyeket a tényleges rövidítési folyamat során ebben a sorrendben kell alkalmaznunk. Pontosan ezen a módon jön létre a r(*Serint*)=*St* – r(*Ser*)=*Sr* szabálypár is (vö: 2(b) ábra).

Fontos, hogy a továbbrövidítés során csak *teljes* rövidítéseket rövidítsünk tovább: r(*ett*)=*eT* esetén ne akarjuk például a *melleT* szóban lévő *elle* elemet rövidíteni, mely egy rövidítéssel részben fed át. Ez ahhoz vezetne, hogy nem tudunk egyértelmű, független szabályokat kialakítani a fenti módon, és több menetben kellene kifejteni a rövidítéseket, ami nagyon megnehezítené a rövidírás olvasását. Továbbrövidítéskor tehát teljes egészében tartalmaznia kell az új szabálybaloldalnak a korábbi rövidítést, azaz a rövidítéseket egy egységként kell kezelnünk. Ezt technikailag úgy oldottuk meg, hogy a (többkarakteres) rövidítéseket megjelöltük egy speciális kezdő- (B) és végjellel (E). A lényeges pont az, hogy az n-gram gyakorisági lista származtatásakor (algoritmus 2. pont) a B..E közötti szakaszon nem vágunk, ezt a szakaszt „egy karakternek" tekintjük. A B*Sr*E*int* elemből képzett n-gramok így a következők lesznek: B*Sr*E, B*Sr*E*i*, B*Sr*E*in*, B*Sr*E*int*, *i*, *in*, *int*, *n*, *nt*, *t*. Az n-gramok hossz-számításakor természetesen a két speciális jelet figyelmen kívül kell hagyni.

Felmerülhet az olvasóban, hogy a rk() számítására bemutatott képlet hiányos. Abban az esetben, ha a rövidítés (jelentős számban) előfordul magyar nyelvű szövegben, azaz a rövidítéssel formailag megegyező eredeti szövegelemek elé a 2.2. részben írtak szerint védőjelet kellene tenni, hogy ne rövidítésként értelmeződjenek. Emiatt a rövidítési képesség csökkenne, a képlet kiegészülne az alább látható utolsó taggal:

$$\mathrm{rk}(w, r(w)) = [l(w) - l(r(w))] \cdot \mathrm{fq}(w) - \mathrm{fq}(r(w))$$

Ezt az utolsó tagot azonban – két okból – elhanyagoljuk. Egyrészt mert végül kizárólag kétjelű rövidítéseket tartalmaz az új szabályrendszer és ezekhez minden esetben találtunk olyan rövidítést, ami egyébként magyar szövegben nem vagy csak nagyon ritkán fordul elő, így védési igénye minimális; másrészt azért, hogy a „mit rövidítsünk" és a „mire rövidítsünk" kérdését valóban szétválaszthassuk egymástól.

Az algoritmusról szóló eszmefuttatás végén megjegyezzük, hogy a bemutatott „mohó" megközelítésről – miszerint ha egyszer kitaláltunk egy szabályt,

akkor azon többet nem változtatunk, sőt le is futtatjuk a teljes szógyakorisági listán, mielőtt továbblépnénk – nem bizonyított, hogy ténylegesen a maximális rövidítési képességű szabályrendszert eredményezi. Azonban mivel komolyan figyelembe kell vennünk a használhatósági szempontokat, és ennek következtében számos, a rk()-t befolyásoló manuális döntést hozunk, esetlegesen elfogadhatjuk a szuboptimális megoldást is kiindulópontként.

6. A korpusz és a rendszer futtatása

Eredetileg a Magyar Nemzeti Szövegtár [10] gyakorisági listájából terveztünk kiindulni. Péter Zsigmond (l. a Köszönetnyilvánítást) javaslatára, hogy a szöveganyagot jobban közelítsük a „vakos" nyelvezethez, végül jelentős mennyiségű ilyen szöveget is hozzávettünk. A *Vakok Világa* folyóirat 31 számának 180000 szónyi anyagát kombináltuk a Szövegtárral 4:1 arányú súlyozással a Szövegtár javára.

A futtatás során legelső lépésként alkalmazzuk a kis rövidírás (2.2. rész) szabályait. Ezeket a rövidítéseket olyan védelemmel látjuk el, ami biztosítja, hogy az eredeti, ismert formájukban megmaradjanak, ne rövidüljenek tovább. Maga az algoritmus tehát már eleve a kis rövidírással rövidített anyag alapján számított gyakorisági adatokat kapja meg.

Egy 60 szabályból álló rendszer előállításakor a futási idő nagyjából 10 perc. Ezt az elfogadható teljesítményt úgy érjük el, hogy a szógyakorisági listának csak az első 50000 bejegyzését vesszük. Ez a szelet a korpusz anyagának 85%-át tartalmazza, így nem torzítja jelentősen az n-gram gyakorisági adatokat, ugyanakkor a futási időt két nagyságrenddel (kb. századára) csökkenti.

7. A rövidítésjelek manuális kiválasztása a használhatósági megfontolások alapján

A használhatósági feltételeknek nagyon nehéz lenne teljesen automatizált módon megfelelni [4]. Szükséges ezért az automatikusan kialakított szabályrendszer interaktív módosítása, manuális véglegesítése szakértők bevonásával a használhatóság maximalizálása érdekében. Szigorúan automatikus úton történik tehát a fenti algoritmussal (5. rész) az épp aktuális (következő) legjobban rövidíthető elem meghatározása, illetve egy ajánlott, jó olvashatóságú rövidítést is automatikusan megad hozzá a rendszer. Ez manuálisan felülbírálható az alábbiak szerint.

7.1. Használhatósági követelmények

A 3. részben említett használhatósági követelmények közül a 2-4. ponttal foglalkozunk most részletesen.

A *jó olvashatóság* azt jelenti, hogy a rövidítések emlékeztessék az olvasót a rövidített szóra. Gyorsolvasáskor gyakran csak a szó első egy-két vagy utolsó

egy-két betűjét olvassuk el, fontos, hogy ezek a betűk az olvasó eszébe juttassák az egész szót. Tapasztalat szerint a jól olvasható rövidítés pozíciótól függetlenül mindig azonos jelentésű, a szó kezdő és záró betűjéből, illetve a szót alkotó jellegzetes mássalhangzóból áll. Ideális eset, mikor teljes szót/szóalakot rövidítünk (nem szórészletet), és a rövidítés a kezdő és a záró mássalhangzóból áll, ahogy ez a kis rövidírásban sok helyen látható: r(*mint*)=*mt*, r(*rövid*)=*rd*. Kiegészítő lehetőség, hogy adott esetben az is megfelelő, ha a jel *kinézete* emlékeztet arra a dologra, amire referál. Agglutináló nyelv lévén magyarban – néhány esettől (pl.: r(*hoG*)=*h*) eltekintve – nem tehetjük meg azt, hogy kizárólag teljes szavakat rövidítünk, mert ez nagyon alacsony rk()-t eredményezne. Ehelyett morfémákat (töveket és toldalékokat) rövidítünk, sőt bizonyos esetekben akár nagyon gyakori szótagokat, és ezeket egymás után illesztve kapjuk meg a szóalakokat. Az itt körvonalazódó elv úgy is megfogalmazható, hogy *„értelmest értelmesre”* rövidítsünk. Azaz lehetőleg értelmezhető elem legyen a rövidítendő, a használt rövidítésből pedig könnyen kikövetkeztethető legyen az eredeti elem.

A *jó felismerhetőség* követelményének akkor felelünk meg, ha tapintás útján könnyen azonosítható jeleket alkalmazunk a rövidítésekben. A 209. oldalon található 1. táblázatban szürkével megjelölt erős jelek felelnek meg ennek a követelménynek. Törekszünk rá, hogy minden rövidítés tartalmazzon erős jelet.

A *könnyű megtanulhatóság* azt jelenti, hogy a szabályok egyszerűek és jellegükben hasonlóak a kis rövidírásban meglévőkhöz, valamint, hogy kevés új szabályt hozunk létre. Az egyszerűség érdekében környezetfüggetlen szabályokat alkalmazunk, az új szabályok számát alacsonyan tartjuk. Ezt minden további nélkül megtehetjük, mert az architektúra lehetővé teszi, hogy a szabályrendszert a jövőben könnyen bővíthessük, ahogy erről a következő részben szót ejtünk.

7.2. A rövidítésjelek kiválasztása

A fenti megfontolások alapján az itt részletezendő kézi módosításokat végezzük az algoritmus eredményeként adódó rövidítendő elem – ajánlott rövidítés párokon. Lényegében minden egyes párról egyedileg döntünk, és utána futtatjuk tovább az algoritmust. Más szóval egyesével vesszük hozzá az új szabályokat a már meglévő szabályrendszerhez. Ez vonja magával azt a lehetőséget, hogy a most kialakított, kevés szabályt tartalmazó javaslat a jövőben bármikor ezen a módon tovább bővíthető igény szerint.

Sok esetben *felülbíráljuk* a rendszer által ajánlott rövidítést, amit a már felhasznált rövidítésjelek ismeretében a rövidítendő elem karaktereiből állít össze heurisztikák alapján. A javaslatban csak kétjelű rövidítéseket használunk. A legtöbb esetben könnyű kiválasztani egy olyan ritka kétjelű rövidítést, ami megfelelően illeszkedik a rövidítendőhöz, pl.: r([maGar])=[mG]. A rövidítésjel gyakoriságát a védési igény minimalizálása érdekében minden esetben ellenőrizzük és egy előre meghatározott küszöb (800/millió szó) alatt tartjuk. Korábban rövidített (szóvégi) elem hangrendi párjához törekszünk ugyanazt a rövidítést rendelni. Ilyen a javaslatban a *-ság/-ség* és a *-nak/-nek*.

Dönthetünk úgy, hogy az adott elemet csak bizonyos *pozícióban* (csak szó elején vagy végén) rövidítjük. Általában azért tesszük ezt, mert lényegében

csak abban a pozícióban fordul elő az adott rövidítendő. Törekszünk rá, hogy az előfordulási arány (például r([meg])=[mg] szó elején) lehetőleg 90% fölötti legyen, hogy ne sérüljön az adott szabály, mint aktuálisan legjobb szabály létjogosultsága. Ilyenkor lehetőségünk van arra, hogy nem általában ritka, hanem csak az adott pozícióban ritka, azaz *komplementer eloszlású* elemet válasszunk rövidítésnek, amint ez a r([lehet])=[lt] esetében meg is történt: a *lehet* elemet csak szó eleji helyzetben rövidítjük, a hozzá társított *lt* rövidítésjel nagyon gyakori ugyan, de lényegében sosem fordul elő szó elején.

Szükség esetén azt is kiköthetjük, hogy az adott elemet – tudva, hogy ezzel veszítünk a rövidítési képességből – *kizárjuk* a rövidíthető elemek köréből használhatósági megfontolások miatt. A javaslat készítése során a következő elemeket zártuk ki: *ala*, *ált*, *áro*, *eGe*, *eke*, *eket*, *ele*, *ere*, *erül*, *ete*, *hat*, *kez*, *kor*, *lat*, *leg*, *let*, *tal*, *tás*, *tele*, *ter*, *tés*, *val*. A puszta betűsorozatok mellett értelmes elemeket is látunk a listán. Ezek főként azért maradtak ki, mert túl ritkák a „kívánt" pozícióban: a *kor* szó végi és a *leg* szó eleji gyakorisága egyaránt 50% alatt marad.

7.3. Elvetett ötletek

A munka során számos ötletet, mely tovább növelte volna a rövidítési képességet, a használhatósági követelmények miatt elvetettünk. Ezek legtöbbször a 7.1. részben említett „értelmest értelmesre" elvet szegik meg, viszont rövidítési képességük jelentős (lenne). Érdemesnek tartjuk, hogy ezekről is említést tegyünk.

Az 5. részben írtak szerint magyarban a legnagyobb rk()-gel rendelkező karaktersorozat az *et*. Értelmetlen karaktersorozatokat azonban a nehéz olvashatóság miatt nem rövidítünk.

Annak ellenére, hogy a kis rövidírásban vannak nagyon hatékony szóközt elnyelő szabályok – a névelőket és a vesszőt érintő szabályokról van szó – ilyen szabályt sem alkalmazunk. A leggyakoribb szóvégi karakter (t) és az azt követő szóköz rövidítése kiemelkedően hatékony szabály: rk([t␣], [T])=1,2%. Az efféle szabályok nyilván rontják az olvashatóságot, mivel több szót egy hosszú egységgé vonnak össze, mégis esetleg érdemes volna megfontolni például a jelenleg *h*-ként rövidített *hoG* kettősponttal való rövidítését mindkét szóköz eltüntetésével: r([␣hoG␣])=[:]; vagy a határozatlan névelő szóközelnyelő rövidítését: r([eG␣])=[;]. E két szabály együttes rövidítési képessége 0,4%.

Mivel egyjelű rövidítéssel nem lehet megfelelni a jó olvashatóság fent leírt követelményének, egyjelű rövidítést egyáltalán nem alkalmazunk. Ugyanakkor a leghatékonyabb szabályok éppen a nagyon gyakori betűkapcsolatok egy karakterre való rövidítései lennének, ezek a szabályok kiemelten értékesek a rövidítési képesség szempontjából. Negyven szabállyal, mely *négy* darab ilyen rövidítést tartalmaz (*et*, *el*, *en*, *er*), közel 17%-os összesített rövidítési képesség érhető el, ami nagyon nagy mértékben meghaladja javaslatunk hatékonyságát. Felmerülhet például a következő elemek egyjelű rövidítése: *el*, *tt*, *meg*, *Ser*, *ás/és*, *eG*, az alábbi nagyon ritka jelek – közülük is főként az erős jelek – bevetésével: [H], [F], [@], [q], [=], [*], [T], [w]. Megjegyzendő, hogy az angol és a német

rövidírásban is alkalmaznak ilyen szabályokat, például r(⠊⠞[it])=⠭[x] az angolban (csak önálló szóként) [1, 103. oldal], r(⠍⠍[mm])=⠭[x] a németben [3]. Hasonló indokok miatt prefixet/posztfixet tartalmazó rövidítést sem használunk.

Ahogy említettük, az egyszerűség érdekében nem használunk környezetfüggő szabályokat. Ez abban nyilvánul meg, hogy a rövidítéseket a rövidítendő környezetétől függetlenül minden esetben alkalmazzuk, valamint, hogy a rövidítések szintén környezetüktől függetlenül mindig ugyanazt jelentik. Ez a döntés az algoritmus egyszerűsítését hozta magával. Amint láttuk, azt megengedjük, hogy bizonyos szabályok csak szó elején/végén legyenek alkalmazhatók, olyan azonban már nem fordul elő, hogy egy rövidítésjelnek két különböző jelentése legyen két különböző pozícióban. A német rövidírásban erre is van példa, érvényes például a következő szabály: r(⠊⠍⠍⠑⠗[immer])=⠭[x] [3], függetlenül attól, hogy az x a (az önálló szóként elő nem forduló) *mm* betűkapcsolat rövidítésére is használatos.

8. A javaslat

A magyar Braille-rövidírás megújítására vonatkozó javaslat a 3. táblázatban látható. A 33 szabály az 5. részben ismertetett algoritmus lefuttatásával, a kimenet 7. rész szerinti manuális finomhangolásával jött létre, megfelelő kompromisszumot alakítva ki a minél nagyobb rövidítési képesség és a használhatóság szempontjai között.

9. A javaslat kiértékelése

9.1. A rövidítési képesség mérésének módszertana

A dolgozatban korábban említett empirikus méréseket, és a javaslat teljesítményének kiértékelését is az itt ismertetendő módon végeztük. A 4. táblázatban látható három tesztfájllal dolgoztunk.

Minden szöveget először a 2.1. részben ismertetett „egy hang egy karakter" formára hoztunk. Megállapítottuk az adott tesztfájl karakterszámát, alkalmaztuk először a kis rövidírás szabályrendszerét, majd az épp mérendő szabályrendszert, meghatároztuk, hogy hány százalékkal csökkent a karakterszám, majd a három tesztfájlra kapott értéket átlagoltuk. A szóközök és az újsor-karakterek is beleszámítottak a karakterszámba. A 2.2. részben írtaktól kis mértékben eltérve a kis rövidírás egyjelű szórövidítéseit két kivétellel kizárólag önálló szóként alkalmaztuk, a *Cak*-ot, az *után*-t és a kétjelű szórövidítéseket viszont minden előfordulásukban. A kis rövidírás esetén nem alkalmaztuk a védőjelet, mivel azt tapasztaltuk, hogy lényegében nincs használatban; az új szabályok esetén viszont minden esetben alkalmaztuk, a kapott értékek tehát az új szabályok védési igényével csökkentett értékek. A mérés során minden rövidítést karakterszintű védelemmel láttunk el, megakadályozva ezzel, hogy rövidítés részletét véletlenül továbbrövidítsük, kivéve persze az egyetlen tényleges továbbrövidítést, a r(*Serint*)=*St* – r(*Ser*)=*Sr* esetét.

3. táblázat. A magyar Braille-rövidírás megújítására vonatkozó javaslat avagy az új magyar Braille-rövidírás. A szabályok előállítási sorrendben vannak feltüntetve. Jelöljük, ha a rövidítést a jel kinézete miatt választottuk, valamint ha csak bizonyos pozícióban rövidítjük az adott elemet. A komplementer eloszlás fogalmát (8., 21. és 32. szabály) a 7.2. részben vezetjük be. A csúsztatott jel fogalma a 2.1. részben található.

	rövidítendő	rövidítés	megjegyzés
1.	 [meg]	[mg]	szó elején (98%)
2.	 [ett]	[eT]	= ott
3.	 [Ser]	[Sr]	
4.	 [ott]	 [oT]	= ett 
5.	 [maGar]	 [mG]	
6.	 [jelen]	[jn]	
7.	 [ség]	 [sg]	= ság
8.	 [lehet]	 [lt]	szó elején (97%) *lt* komplementer: gyakori, de nem szó elején
9.	 [vezet]	[vz]	
10.	 [nek]	[nx]	= nak ; szó végén (83%)
11.	 [köz]	[kz]	
12.	 [nak]	 [nx]	= nek  ; szó végén (96%)
13.	 [fel]	 [fl]	szó elején (86%)
14.	[Serint]	[St]	szó elején (95%)
15.	 [rend]	 [rH]	csúsztatott *d*  *rd* foglalt a kis rövidírásban (*rövid*), *rn* túl gyakori
16.	 [ember]	 [wr]	jel kinézete
17.	 [kormáN]	 [kN]	
18.	 [ellen]	[oó]	jel kinézete
19.	[elnök]	[eö]	
20.	[elő]	[eő]	
21.	[tart]	[tt]	szó elején (78%) *tt* komplementer: gyakori, nem szó elején
22.	[áll]	[Ll]	jel kinézete (egyéb ötlet: [áy])
23.	[mond]	[mH]	csúsztatott *d* *md* foglalt a kis rövidírásban (*mind*)
24.	[támogat]	[tg]	
25.	[ért]	[éT]	
26.	[ság]	[sg]	= ség
27.	[következ]	[kv]	
28.	[akkor]	[ao]	
29.	[budapest]	[bp]	
30.	[kerül]	[eü]	
31.	[történ]	[öé]	(egyéb ötlet: [Tn])
32.	[több]	[tb]	szó elején (95%) *tb* komplementer: gyakori, de nem szó elején (egyéb ötlet: [t=])
33.	[kapColat]	[kC]	

4. táblázat. A rövidítési képesség méréséhez használt tesztfájlok.

megnevezés	méret
egy zenei híreket tartalmazó fájl az MVGYOSZ-ből	11000 szó
a *Vakok Világa* 31 számának anyaga	180000 szó
Mikszáth Kálmán: *Szent Péter esernyője* c. regénye	53000 szó

9.2. A kiértékelés eredménye

A 8. részben ismertetett javaslatunk kiértékelésének eredménye az 5. táblázatban látható. A javaslat kevés szabályt tartalmaz, könnyen megtanulható. A rövidített szöveg olvashatósága kiváló, felismerhetősége is megfelelő. A rendszer rövidítési képessége kielégítő: a kis rövidírás által képviselt rövidítési képességet harmadával megnöveltük.

A 13,3%-os eredmény az angol Braille-rövidírás közel 20%-os rövidítési képességével [11] összevetve nem tűnik soknak. Érdemes ugyanakkor látni, hogy az angol rendszer majd 200 szabályt tartalmaz, és számos, a 7.3. részben leírt eljárást kiterjedten alkalmaz. Nehezebben tanulható, olvashatósága jelentősen rosszabb. Javaslatunkban az egyes szabályok átlagos rövidítési képessége 0,1%. Ezt is érdemes összevetni a 7.3. részben említett példák rövidítési képességével.

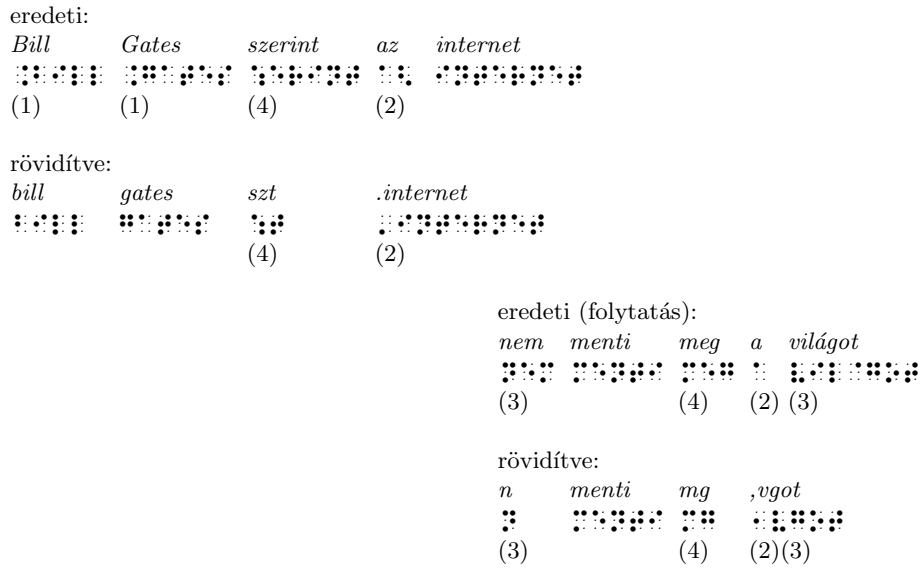
5. táblázat. A javaslat 33 szabályának összesített rövidítési képessége.

tesztanyag	kis rövidírás rk()	új rövidírás rk()	Δrk()	Δrk() %
zenei hírek	9,5%	12,5%	+3,0%	
Vakok Világa	9,5%	13,9%	+4,4%	
Szent Péter...	10,7%	13,5%	+2,8%	
átlag	**9,9%**	**13,3%**	**+3,4%**	**+34%**

Az új rövidítendő elemek ábécérendes listája a következő: *akkor*, *áll*, *budapest*, *ellen*, *elnök*, *elő*, *ember*, *ért*, *ett*, *fel(-)*, *jelen*, *kapColat*, *kerül*, *kormáN*, *következ*, *köz*, *lehet(-)*, *maGar*, *meg(-)*, *mond*, *-nak/-nek*, *ott*, *rend*, *ság/ség*, *Ser*, *Serint(-)*, *támogat*, *tart(-)*, *több(-)*, *történ*, *vezet*

9.3. Példák

A 3. ábrán egy példamondaton mutatjuk be az új rövidírás működését, a 6. táblázatban pedig jellegzetes új rövidítéseket tartalmazó szavak láthatók.



3. ábra. Az új rövidírás működése egy példán. A kis rövidírás (1) elhagyja a nagybetűjelet; (2) összevonja és rövidíti a határozott névelőket: (3) szórövidítéseket alkalmaz (*nem*, *világ*). Ez után következnek (4) az új rövidítő szabályok (*Serint*, *meg*). A mondat hossza 55 karakterről 40 karaterre csökkent, ami 27,3%-os rövidülést jelent.

6. táblázat. Jellegzetes új rövidítéseket tartalmazó szavak.

rövidítendő	rövidítés	Δhossz		
gyakran előforduló rövidített szavak				
[között]	[között]			
[eGSer]	[eGSr]			
[lehetett]	[lTeT]			
nagy arányban rövidülő szavak				
[kapColatban]	[kCb]	-8	72%	
[következő]	[kvő]	-6	66%	
[maGarorSág]	[mGog]	-6	60%	
három rövidítést tartalmazó, sok karakterrel rövidülő szavak				
[kapColatrendSer]	[kCrHSr]	-9	60%	3új
[jelentőségteljes]	[jntősgts]	-8	50%	1r+2új
[boldogságban]	[bgsgb]	-7	58%	2r+1új

10. Konklúzió

Az eredeti alapötlet bevált, a közel maximális rövidítési képességgel bíró, ugyanakkor kényelmesen használható új magyar rövidírás szabálykészlete félautomatikus módon, a rövidítendők korpuszvezérelt meghatározásával és a rövidítések kézi finomításával előállítható.

Az összesített rövidítési képesség 13,3%. Ez jelentős – több mint 30 százalékos – növekedés a kis rövidítés 9,9%-os hatékonyságához képest, a használhatósági követelmények miatt azonban ez jóval alacsonyabb, mint a lehetséges elvi maximum. A módszerből adódóan a szabályrendszer a jövőben könnyen bővíthető. Jóval több és/vagy nehezebben olvasható szabállyal természetesen megközelíthető akár a 20% is. Esetünkben az volt a koncepció, hogy az új magyar rövidírás bevezetésének megkönnyítése érdekében a változtatás, bővítés mértékével nagyon óvatosak voltunk, és a kompromisszumos javaslat kialakítása során nagyobb hangsúlyt fektettünk a használhatóságra, mint a rövidítési képesség minden áron való növelésére. A rövidítendő elemeket meghatározó korpuszvezérelt algoritmusnak köszönhetően az adott feltételek mellett az objektíve legjobb rendszert hoztuk létre.

Ha összevetjük a kis rövidírásban lévő kétjelű rövidítések (210. oldal: 0,05%), illetve az új kétjelű szabályok egy szabályra eső rövidítési képességét (220. oldal: 0,1%), akkor azt látjuk, hogy a korpuszvezérelt módon létrehozott rendszer még úgy is *kétszeres* teljesítményre képes az intuíció, illetve hagyomány talaján álló rendszerrel szemben, hogy már eleve jelentősen rövidített szövegen kell dolgoznia.

Konklúzióként levonhatjuk tehát – és ez érvényes lehet a különféle annotációs vagy ontológiaépítési feladatoktól, a szótárkészítésen át akár egyes elméleti nyelvészeti kérdésekre is –, hogy ha valamit meglévő (gyakorisági) adatokból automatikusan származtatni tudunk, akkor nem érdemes intuitív megközelítést alkalmazni. Vagy másképp fogalmazva: érdemes az intuíciót bizonyos adatvezérelt módszerekkel legalábbis kordában tartani.

A rendszer alkalmas a bevezetést előkészítő közvetlen vakok általi tesztelésre, bízom benne, hogy találkozhatunk majd vele Braille-nyomtatványokban vagy akár a közterek Braille-feliratain.

Köszönetnyilvánítás

A projekt munkálatai kapcsán nagy köszönettel tartozom a Magyar Vakok és Gyengénlátók Országos Szövetsége Braille-bizottsága részéről *Péter Zsigmond*-nak, aki a Braille-írás és a Braille-rövidírások szakértő ismerőjeként a munka számos pontján volt segítségemre: megismertetett a Braille-írással, ellátott szakirodalommal, bevezetett a használhatósági megfontolások rejtelmeibe, a javaslat kialakítása során pedig közösen hozhattuk meg a rendszert érintő konkrét használhatósági döntéseket.

Hivatkozások

1. Simpson, Ch., ed.: The Rules of Unified English Braille. Version I. Round Table on Information Access for People with Print Disabilities Inc., Australia (2010)
2. Freud, E.: Leitfaden der deutschen Blindenkurzschrift: Teil 2. Verlag der Deutschen Blindenstudienanstalt, Marburg (1973)
3. `fakoo.de`: Einführung in die deutsche Braille-Kurzschrift `http://www.fakoo.de/kurzbraille.html`
4. Sass, B.: Az új magyar Braille-rövidírás korpuszvezérelt kialakításának lehetőségei. In: IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2013), SZTE, Szeged (2012) 348–350
5. Flamich, M., Hoffmann, R.: A tapintható írásrendszerek történeti áttekintése. Iskolakultúra **20**(1) (2010) 3–17
6. Görgényi, M., ed.: A magyar pontírás. Teljesírás. Pécs (1998)
7. Görgényi, M., ed.: A magyar pontírás. Rövidírás. MVGYOSZ, Budapest (2001)
8. Bogart, D.: Unifying the English Braille Code. Journal of Visual Impairment & Blindness **103**(10) (2009) 581–583
9. Arató, A.: A BraiLab beszélő számítógépcsalád. Kandidátusi értekezés. (1992)
10. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002), Las Palmas, Spain (2002) 385–389
11. Durre, I.K.: How much space does Grade 2 Braille really save? Journal of Visual Impairment & Blindness **90**(3) (1996) 247–251